DOCUMENT RESUME

ED 279 666 TM 870 053

AUTHOR Carroll, John B.
TITLE Scales and Other Problems in the NAEP Reading Assessment: Critical Comments.
PUB DATE 18 Aug 86
NOTE 41p.; One of 46 papers commissioned by the Study Group on the National Assessment of Student Achievement and cited in Appendix B to their final report "The Nation's Report Card" (TM 870 049). For other papers in this group, see TM 870 050-094.
PUB TYPE Viewpoints (120) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Assessment; Elementary Secondary Education; Equated Scores; *National Surveys; Rating Scales; Reading Achievement; *Reading Skills; *Reading Tests; Scaling; Scores; Test Construction; *Testing Problems; *Testing Programs; *Test Interpretation; Test Results; Young Adults
IDENTIFIERS *National Assessment of Educational Progress; Sequential Tests of Educational Progress; Unidimensional Scaling

ABSTRACT

Two major issues are addressed in this commentary on the National Assessment of Educational Progress (NAEP) reading assessment: the design of the reading proficiency scale and interpretation of test data; and the dimensionality of this scale and the possibility of supplying data on specific skill dimensions. An examination of NAEP data and the five levels indicated by the scale scores indicated problems in interpretation and reporting of test results. Controversies concerning the validity and measurement of reading subskills are discussed. Although the current NAEP reading scales are considered acceptable, several recommendations are made: (1) supplementary information on readability levels and grade placements should be provided for the Reading Proficiency Scale; (2) reports furnished to the public should express the variability features of assessment data; (3) proficiency in word decoding, vocabulary knowledge, and reading speed should be measured; and (4) although testing need not be extended to children under 9, testing of a representative sample of adults is recommended. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED279666

Scales and Other Problems in the NAEP Reading Assessment:

Critical Comments

John B. Carroll

University of North Carolina at Chapel Hill

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☑ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Carroll

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper commissioned by

THE STUDY GROUP ON THE NATIONAL ASSESSMENT OF STUDENT ACHIEVEMENT

1986

TM 870 053.

BEST COPY AVAILABLE

Scales and Other Problems in the NAEP Reading Assessment: Critical Comments

John B. Carroll

University of North Carolina at Chapel Hill

August, 1986

[Prepared for the Subcommittee on NAEP Reading Assessment, chaired by Jeanne Chall]

These comments are based mainly on information contained in The Reading Report Card (ETS, 1985, hereafter referred to as the Report), but also on supplementary information kindly supplied by Dr. Albert E. Beaton of Educational Testing Service, who was one of those principally responsible for the design of the Year 15 NAEP Reading Assessment. I will focus on two major issues:

(1) The design of the scale of reading proficiency and the interpretation of the assessment data using this scale.

(2) The dimensionality of this scale and the possibility of supplying data on specific skill dimensions,

I will also make brief comments on several other issues.

I. The Reading Proficiency Scale

ETS is to be commended for its attempt to provide a "common scale that, for the first time, allows direct comparisons across ages" (Report, p. 6). The common scale was developed on the basis of recent advances in psychometrics, chiefly in item response theory (IRT). The Report further states that this scale, which ranges from 0 to 500, "also improves NAEP's ability to make comparisons over time and among subgroups of the population." Five levels of reading proficiency are identified and attached to points on the scale: Rudimentary (150), Basic (200), Intermediate (250), Adept (300), and Advanced (350). The Report gives fairly detailed verbal descriptions of these levels in its Figure 2.3 (p. 15), and sample items purportedly associated with each of the levels are presented on pages 16-28. Some information on how the levels were arrived at and scaled is given in a Procedural Appendix, pp. 62-64, but more detailed information is to be found in a draft technical report that has been made available to me.

The scale is inevitably arbitrary in its metric, and there is no guarantee, unless one accepts certain assumptions about the normality of reading ability distributions, that the intervals on the scale are in some sense equal. The labels and descriptions of selected points on the scale are also inevitably arbitrary; they acquire meaning only in terms of inferences one might make about the levels of reading ability an examinee has to have to perform at some high probability of success the exercises that are placed at the given scale points. From the draft technical report, one

learns that the five levels selected for description correspond, respectively, to the -2, -1, 0, +1, and +2 sigma points on the total distribution of reading proficiency over the age 9, 13, and 17 samples combined. Not unexpectedly, the point 250 ("intermediate" proficiency) is close to (actually a little below) the mean score for the age 13 sample. One might question whether the five points selected for description are optimal, but I see little point in doing so since they seem reasonably appropriate.

It is possible to take either of two opposing points of view about this scale. On the one hand, one might decide that since a considerable amount of thought and effort has already gone into the scale, it is sufficiently satisfactory for further use and therefore should be accepted as it stands. On the other hand, one might decide that the scale has such critical defects that efforts should be directed toward modifying it radically or even replacing it with a new scale or scales. I will not immediately take a stand as to which point of view I think should be favored; first we need to consider what the possible defects of the scale are.

One possible defect in the scale is that the words attached to five scale points (150, 200, 250, 300, and 350) are generally "weasel words" that may lead to confusion and misinterpretation. Despite the verbal descriptions offered for these points and the displays of sample exercises associated with them, people can have different concepts of what "rudimentary," "basic," "intermediate," "adept," and "advanced" skills might be. To me, the terms "intermediate" and "adept" seem particularly unfortunate.

"Intermediate" skills might be at almost any level--"intermediate" between something lower and something higher, and the word "adept" might apply to almost any level of higher skill. One would not want to claim, for example, that a reader far above point 300 on the scale is not an "adept" reader. On the other hand, one could defend NAEP's choice of terms by citing Lewis Carroll's Humpty Dumpty, who was made to say, "When I use a word, it means just what I choose it to mean--neither more nor less." No natural language--certainly not English--is very good at offering terms to describe precise points on a graded scale, and we may have no option but to accept NAEP's terms with the hope that over time they will acquire their intended meanings in the minds of those who use the scale.

It is important to note that the scale assumes a kind of unidimensionality in the development of reading ability; that is, that reading ability, even if it is seen as involving a host of separate, identifiable skills (letter recognition, word decoding, lexical knowledge, syntactic comprehension, inferential reasoning, and use of context and background knowledge, for example), tends to develop along a single dimension, if viewed in the long run. In other words, like any other highly complex skill (e.g., playing and performing on a musical instrument), reading ability develops through a number of phases that on the whole must occur in a definite sequence. Chall (1983) has made proposals about what these stages are. Although her proposed stages do not correspond exactly with the five levels on the NAEP proficiency scale, the NAEP reading scale reflects a recognition of some such stages, in

that the lower reaches of the scale (up to about the "Rudimentary" level, at 150 on the scale) correspond roughly to the stage at which the reader acquires elementary processes of letter and word recognition; the "Basic" level (at 200 on the scale) corresponds to a succeeding stage in which the reader can use elementary skills to get meaning out of simple texts; and so on. Thus, the hope is that with the use of the NAEP scale it is possible to index where a reader is on the continuum of reading skill thus conceived. Use of the scale implies a kind of "criterion referencing" of reading performances, and from this standpoint the scale has much appeal both in terms of psychometric adequacy and in terms of practical utility and ready understanding.

There are, however, at least two aspects of reading skill that are, so to speak, concealed or latent in the NAEP reading scale.

One is the concept of reading difficulty or "readability" of text. There is a long tradition of research on readability (Carver, 1974; Chall, 1953; Klare, 1963); the concept, at least, is reasonably well defined. Although there are numerous methods of measuring it, its measurements are generally substantially correlated, even if not correlated perfectly. From an inspection of the exercises given as exemplars of points on the reading scale, it becomes evident that the scale might be used as a fairly accurate description of the levels of readability of the texts used in these exercises. The report does not indicate anything about these levels of readability; it would have been helpful if

it had done so. If preparers of the report were hesitant about this because of the problem of selecting a measure of readability, several well-accepted measures could have been used along with appropriate explanations. In any case, it would help to define points on the scale if associated indices of readability could be attached to them.

The other latent variable in the NAEP reading scale is grade placement. In view of the many controversial questions about grade placement, one can perhaps understand the reluctance of the NAEP researchers at ETS to use grade placement indices, but the fact is that grade placement is a concept that is readily understandable by teachers, parents, and even students. That is, teachers, parents, and students understand that in the normal operation of the American school system at any given epoch, there are distinct levels of reading ability that are expected to be attained by the _average_ student in each grade. The _Report_ makes little reference to grade (an exception: "9-year-olds are primarily in third and fourth grade", p. 17), and given the data of the report it is impossible to relate the reading scale to grade placement with any certainty. We are not even told the means and variances of the grades in which the 9-, 13-, and 17-year-old samples were found. (From information in the draft technical report, the mean grades for the three samples are found to be about 3.7, 7.7, and 10.9, respectively.) I feel, in any case, that much would be gained if the NAEP report were to report information about grade placements -- even if only in the form of quite factual information about the median or modal grade

placement of students attaining different levels on the NAEP reading scale. This information could of course change from one assessment to another, and would be useful in evaluating trends in levels of reading ability either nationwide or for specific regions or groups.

In view of the fact that information about readability levels and grade placements is not included in descriptions of reading proficiency levels, some doubt can be expressed about the real meaning of those levels and the meaningfulness of the verbal descriptions given in the Report's Figure 2.3. One can imagine the frustration experienced by the experts asked to construct those verbal descriptions from data on exercise performance. Consider, for example, the statements included for the Basic (200) and Intermediate (250) levels: "Performance at [the Basic] level suggests the ability to understand specific or sequentially related information." "Performance at [the Intermediate] level suggests the ability to search for specific information, interrelate ideas, and make generalizations." It is unlikely that these statements are adequately specific to these levels. Ability to understand specific or sequentially related information could apply to all levels; success would depend on the readability of the texts, vocabulary level, syntactic complexity, etc., as well as the types of questions asked. Likewise, the ability to search for specific information, interrelate ideas, and make generalizations could apply at almost any level of reading proficiency, with success depending on readability and other difficulty factors. For this reason, the verbal descriptions need

to be supplemented with objective data on passage readability and grade placement.

The report could have done much more in indicating the meaning of levels on the NAEP reading scale in terms of success in performance of typical exercises. It is not clear exactly what level of performance on exercises is indicated by a given value on the scale. One is told only that

> In the scale anchoring process, NAEP selected sets of items that were good discriminators between proficiency levels. The criterion used to identify such items was that students at any given level would have at least an 80 percent probability of success with these reading tasks, while the students at the next lower level would have less than a 50 percent probability of success [on these same tasks]. (Report, pp. 63-64)

To say that a probability of success was "at least 80 percent" or "less than 50 percent" does not give a sufficient indication of the average probabilities of success attained by students at given points on the NAEP reading scale with items at different difficulty levels. There is a simple way of exhibiting such probabilities, and I am sure that NAEP could develop charts and graphs from its available data that would be much more meaningful, using what I (Carroll, 1985) have called the person characteristic function. This function plots probability of success against item difficulties for groups at different points on the reading scale. Figure 1 gives a hypothetical example of such a plot; I am unable to make it much more than hypothetical because relevant data are not available in the Report. That is, I am unable to tell what

liminal level of success was used in placing the items on the scale. (In plotting the figure, I have assumed a probability of 90%, and a probability of 35% for success at the next higher scale point at an interval of 50. I have also assumed a ζ (guessing) parameter of .2, which is very near the median value of ζ for the 228 items scaled by BILOG, according to information supplied to me).

Insert Figure 1 about here

In using this procedure, it is probably desirable to group items of similar difficulty levels and plot the average probabilities of success by groups of examinees whose average reading scores correspond to selected points on the reading scale. One advantage of the procedure is that it reports, for any selected group of examinees, the probabilities of success for items at different points on the scale, and gives an idea of the discrimination values of the items.

Accepting the scale as it stands, I now want to consider how well it serves in communicating the results of the four NAEP reading assessments that have been reported in The Reading Report Card. Actually, I am more concerned with the method of reporting the results, which I think is somewhat inadequate and, to a certain extent, misleading because it largely ignores variability and what has happened over the years in different regions of the spectrum of reading ability. It is only through some statistical processing, with certain assumptions, of data given in Table 2.1 (Report, p. 16) that I have been able to estimate more clearly the

variability features of the data.

My methodology was as follows: I first made the assumption that total reading ability as indicated by the NAEP scale was normally distributed at each assessment, for each age group. This assumption is of course questionable but I had no alternative, and the assumption is at least reasonable as yielding approximations to the actual distributions. (NAEP could easily compile data from the actual distributions; also, it appears from graphs included in the draft technical report that the distributions are indeed approximately normal.) The percentage values in Table 2.1 of the Report were converted to their normal deviate equivalents and then plotted against the NAEP scale for each age group, for a given assessment. (For convenience, only the 1971 and 1984 data were used.) These plots were approximately linear; in some instances, lines were extrapolated to estimate certain percentile points. From these plots, certain percentile points for the NAEP scale were estimated, namely, the 0.5, 5, 25, 50, 75, 95, and 99.5 percentile points for each age group. These percentile points were selected to correspond to those used in my presentation (Carroll, 1971, Figure 3-1) of results (Educational Testing Service, 1957) from the STEP Reading Test norms, which purportedly were based on a scale similar to the NAEP scale in the sense that it spans nearly the total spectrum of reading ability. For convenience, my previous figure is reproduced here as Figure 2.

Insert Figure 2 about here

The NAEP scale values were then plotted against age for each percentile point, yielding Figure 3 for the 1971 assessment and Figure 4 for the 1984 assessment.

Insert Figures 3 and 4 about here

Possibly the reader can work out some way to superimpose these figures. If this is done, it will be seen that there is very little difference between the figures. The 1984 data show a few very slight increases for all ages for the 25, 50, and 75 percentile points--possibly reflected in the "significant" differences reported in the Report's Table 2.1. (I put quotes around the word significant to emphasize that the small differences were probably significant only because the sample sizes were very large.) There is no apparent difference at any age at the 95 percentile point, and an apparently marked decline, for all ages, at the 99.5 percentile point. The only marked increases are for percentile 5, for ages 13 and 17 only, and for percentile 0.5, chiefly for age 17. It thus appears possible that the increases in the means from 1971 to 1984 (as shown in the Report's Figure 2.2) are largely accounted for by the increases at the lower region of the distribution. That is, the increases reflect somewhat better attainment of "rudimentary" and "basic" skills by students in the lower tails of the distributions, and may actually be depressed by declines at the top of the distribution. Any advances in "intermediate" and "adept" skills

in the middle of the distribution, even though they may be statistically significant, are hardly impressive.

What is more, the patterns of results, for both 1971 and 1984, are highly similar to the pattern shown for the Step Reading Test in Figure 2 (the data are for some year prior to 1957). From the data at hand, it was impossible for me to estimate any consistent conversion of the NAEP scale to the STEP Reading Test converted score. There is no way of telling whether average reading ability tested in 1971 or 1984 is better or worse than reading ability measured in the 1950s by the STEP Reading Test, and the STEP data are shown by grade level rather than by age. What all these data show is that there continues to be an enormous variability in reading skills at all three ages studied in the NAEP assessment, with substantial proportions of the samples well below the "intermediate skill" level even at Age 17. This fact somehow does not hit one in the eye in reading The Reading Report Card, and thus tends to be overlooked by educational policy makers (and politicians as well). It could have been pointed out, for example, that about 5% of students in the age 17 sample are below the average of the Age 9 sample.

It would, I think, be useful in a future NAEP assessment to administer the STEP Reading Test to a sample of students that would be sufficient to establish a conversion between the NAEP scale and the converted score on the STEP. (Possibly appropriate data are already at hand somewhere.) This would provide further historical perspective on national progress (or lack thereof) in

literacy education. One can easily assume that NAEP and STEP scores would be correlated highly enough to permit a reliable conversion.

Although it may be unpopular to make the suggestion, I cannot resist commenting that the high variability in reading progress may reflect variability in basic cognitive abilities, of either genetic, constitutional, or environmental origin, that are required for success at more advanced levels of reading skill. For many individuals, low levels of cognitive ability may limit reading success to "rudimentary", "basic", and "intermediate" levels, even under optimal instruction. The results discussed above suggest that nationwide reading efforts may have at least accomplished something for the lower end of the distribution. They have done little for the great middle hump, and have for some reason signally failed for the very top group. But the better results for the lower end of the distribution at Age 17 may actually be illusory, since it seems possible that they result from a dropout effect. The NAEP age 17 results are limited to those age 17 persons still in school; presumably, the skills of those age 17 persons who are no longer in school would be generally low.

The above remarks strictly apply only to the results for total samples. I have not attempted to make similar analyses for the various subgroups whose results are shown later on in the Report, but my conclusions do not seem inconsistent with conclusions presented in the Report for black, Hispanic, and white

students on pages 35-41, for example.

I implied earlier that I would render some judgment about the acceptability of the NAEP scale as it now stands. I feel that on the whole it is acceptable and should be allowed to stand for future assessments, but that several kinds of supplementary information would make it more useful. First, the scale should be indexed in terms of some measure or measures of readability. Second, it should be indexed in terms of expected grade placements, to the extent that this would be possible. Third, more meaning should be associated with further points on the scale; that is, data should be furnished on what kinds of performances would be associated with points 0, 50, and 100, and points 350, 400, 450, and 500. Assigning meanings to the lower points on the scale might be especially necessary if, as may be proposed, assessments were to be extended to lower ages like 5 or 7, for example. Designation of meanings for the higher points would be necessary if assessments were to be made for adult literacy (as may be already under way). I would also strongly recommend use of the person characteristic function to display performance levels on the scale. Finally, further thought and research should be devoted to selection of adequately descriptive terms and phraseology for selected points on the scale.

II. The Dimensionality of Reading Proficiency: Are Specific Scales Needed?

Contained in the draft Technical Report, Vol. 2, on the Year 15 NAEP Assessment is a section by Rebecca Zwick of ETS on the dimensionality of the data from which the Reading Proficiency Scale was derived. After reviewing the results of four somewhat different statistical approaches to this issue, Zwick concluded that "it is not unreasonable to regard the reading items as measures of a single dimension."

For at least two reasons, one need not be surprised at this conclusion.

First, all the methods that were employed to investigate dimensionality are fraught with statistical problems that might have precluded clear determination of meaningful dimensions beyond a single first factor even if they existed. Here I am not questioning the results of Zwick's analysis; rather, I am pointing out the difficulties in coming to any firm conclusions--difficulties that are acknowledged by Zwick.

Second, and more importantly, all the items analyzed were strictly measures of various highly related reading skills; exercises thought to be testing study or reference skills were not included in the analysis. (Nor, in fact, were such items included in the item calibration procedures that resulted in the Reading Proficiency Scale.) Even in one of Zwick's analyses that attempted to maximize the chances of distinguishing scales for

different item types (reading comprehension, vocabulary, life skills, and essay), a second factor was not statistically significant, and the first factor accounted for about 39% of the variance--far greater than accounted for by any further factor. The four item types could all be said to be measures of general reading proficiency, and it is not surprising that only one factor emerged in this analysis. In view of the data analyzed--sets of reading items widely differing in difficulty in samples that were rather heterogeneous in reading ability--it would not be unreasonable to expect that these items would turn out to measure proficiency on a single common scale.

It should also be noted that apparently most or all of the NAEP reading exercises that were analyzed required reading of passages longer than a single word, and thus each may have called on a variety of reading skills. That is, no exercises were really focused on, for example, word decoding apart from other reading skills, or on vocabulary apart from other reading skills. The construction of the exercises was based, of course, on predetermined reading objectives that emphasized global rather than specific skills.

Authorities in the field of reading have always acknowledged that reading ability involves various rather specific skills. To be sure, there has not always been total agreement as to what these skills are, but theoretical developments and research over the past 50 years or so, and particularly in recent years, have produced a general consensus of opinion that recognizes certain

major skills that do not necessarily develop in parallel fashion in all readers. Among such skills are the following:

(1) Word decoding. This is the skill of using the alphabetic principles underlying English spelling patterns as aids in determining what a word is and in determining its probable pronunciation. The skill is most apparent and critical when the word to be decoded is, for example, an unfamiliar or unusual proper name in a telephone list where there is no context to help, but it is also useful even if the reader is trying to identify a word from its context. I forgo any lengthy discussion of this point, despite the enormous amount of frequently misinformed controversy that has surrounded it in the reading instruction literature. A good exposition of the point is to be found in Perfetti's (1985) chapter on beginning reading. Perfetti makes the "central point" that "ensuring that a child learns about the code is a worthwhile instructional objective" (p. 241). The degree to which a reader--at any age--has learned about the code is a worthwhile thing to find out about in a reading assessment program and was in fact recognized as such in the list of reading objectives prepared for the earliest NAEP reading assessment (NAEP, 1970). A separate scale of word decoding is needed (even if it might be substantially correlated with general reading proficiency) so that we can learn the degree to which reading programs in general have been successful in teaching this skill, whatever their success has been in teaching general literacy.

(Incidentally, in searching through my files I came across records of a conference held at ETS June 29-30, 1974, sponsored by

a "Right to Read Committee" whose members included Jeanne Chall. The conference was concerned with establishing a definition of word decoding for the purpose of developing a test of decoding ability in connection with adult literacy. I'd be curious to know what, if anything, ensued as a result of this conference. I mention it not only to emphasize the importance of word decoding at all levels of reading ability but also to point to the possibility that useful work on the problem may already have been done at ETS.)

(2) Vocabulary knowledge. It seems obvious that the extent of vocabulary knowledge a reader possesses is often critical in determining the degree of difficulty in a reading passage that a reader can successfully master. Certainly many of the more difficult exercises in the NAEP series are difficult mainly because they require a high level of vocabulary knowledge. In my opinion it would be worthwhile to single out vocabulary knowledge as a specific dimension of reading proficiency. If a scale of vocabulary knowledge were to be constructed and administered at a future assessment, it would no doubt be found to be highly correlated with general reading proficiency, but not necessarily up to limits imposed by reliability. The implicit or explicit teaching of vocabulary is one concrete and practical component of a balanced program of reading instruction and NAEP information on vocabulary knowledge attained at various ages would be of help in evaluating this aspect of reading instruction in the US.

(3) Reading speed. One component of reading skill that was not directly measured in NAEP assessments that have been conducted up

to 1984 is reading speed. In general, the exercises were administered under certain time limits and thus students were able to complete them successfully only if they had acquired a certain degree of speed in reading. It is entirely possible that many readers could have completed exercises more successfully if time limits had been longer. All the descriptions of points on the Reading Proficiency Scale need to be qualified in terms of the reading speed requirements in attaining them. As the descriptions currently stand (Report, Figure 2.3), nothing is stated concerning speed. It would probably be difficult to analyze current data to arrive at statements about attained levels of reading speed. This deficiency should be remedied in future assessments, probably by developing exercises that would explicitly measure speed in arriving at comprehension.

One could possibly mention still other specific reading skills that might be addressed in future assessments. The three named above appear to me to be the most important, however, and in any case testing more than three separate skills would probably make the NAEP assessment a more complicated process than would be acceptable. (One further skill that might be considered is listening ability, as an index of overall language development independent of reading, but this is not strictly a reading skill and its appropriateness for a reading assessment might be questioned.)

In the year 1977 I was a member of a committee called together by the Education Commission of the States (ECS) to

propose objectives for the NAEP assessment planned for 1979-80. At least some members of the committee, including myself, were strongly in favor of explicitly testing word decoding, vocabulary, and reading speed, and for a time a plan for doing so was accepted by ECS. The plan was ultimately dropped, however, apparently because of objections raised by a professional organization in the reading field. Because it is possible that such objections could be raised again, it is necessary to address and counter them.

According to my notes, the official minutes of the International Reading Association Board of Directors meeting of January 27-29, 1978, recorded the passing of a resolution that read as follows:

> The IRA Board of Directors is concerned about the direction of NAEP in attempting to assess subskills, and strongly recommends that NAEP assess only functional reading ability and not attempt to assess the subskills of reading, referred to as "enabling skills" by the NAEP Advisory Committee, because:
>
> 1. There is no agreement as to what these subskills are.
> 2. There is no agreement as to the hierarchy of the subskills.
> 3. There is evidence that mastery of these subskills does not necessarily lead to reading and, further, that the absence of these skills is not necessarily concomitant with limited reading achievement.
> 4. A test which assesses so-called "enabling skills" will be viewed by many educators as having diagnostic value; yet an instrument which has to serve the broad purposes of NAEP cannot be diagnostic.
> 5. The selection of certain subskills over others will give undue emphasis to a few, perhaps even questionable skills, over the vast array of possible skills.
> 6. The reporting of subskills results may lead to a fractionated, mechanistic approach to the teaching of reading.
>
> For the above reasons, we are convinced that reading

> instruction may be harmed because of the potential misuse of the test results. Responsible test development and use must include concern for the possible deleterious effects of a test on instructional programs.

There may possibly have been a reasonable basis for such a statement in 1977, but there seems to be little basis for it at present, in the light of research and theory development since then.

With regard to the resolution's first point, there is now very substantial agreement about subskills of reading. That is, the conclusions of many prominent researchers in this field are highly convergent; see, for example, work of Baron (1977), Curtis (1980), Frederiksen (1982), Jackson and McClelland (1979), Perfetti (1985), and many others. This work clearly supports the conclusion that word encoding, vocabulary knowledge, and speed of lexical processing are important and critical skills of reading. Thus the first argument made by the 1978 IRA resolution cannot be supported.

The second argument, that there is no agreement as to the hierarchy of reading skills, is vague and imprecise. It is not clear what is meant by a "hierarchy" nor is it clear how any such hierarchy would be relevant to reading instruction or assessment. In any event, no claims are being made here about any such hierarchy among the three skills I am proposing be tested except to the extent of asserting that all are components of general, functional reading skill.

The third point is weak and is inconsistent with point one in

its implied assumption that certain subskills can indeed be identified and agreed upon, as would be necessary in order to cite any "evidence" about the effects or non-effects of their mastery. The resolution probably referred to putative evidence that skill in word decoding does not necessarily lead to functional reading and that some functional reading ability can be attained in the absence of word decoding skill. Even if these propositions are true, as they may well be, they afford no basis for rejecting a plan to assess word decoding skill in a NAEP assessment. If anything, they provide an argument for including this skill in an assessment, because they suggest that word decoding is indeed a skill that is somewhat independent of some aspects of functional literacy. I refer again to the writings of Perfetti (1985) for a discussion of the research evidence surrounding the issue.

Point 4, about the potential interpretation of NAEP data as having "diagnostic" value, is again unclear, since it does not define the term diagnostic. Certainly NAEP data will never be diagnostic in the sense of identifying strengths and weaknesses of individual students. NAEP data could be diagnostic in the sense of pointing to strengths and weaknesses of groups of students at various points on the Reading Proficiency Scale and in this respect the information could certainly be useful. It is hard to imagine how such information could be detrimental to reading instruction.

Point 5 complains that the selection of certain "possibly questionable" subskills over others would be unfortunate, and

claims that there is a "vast array of possible subskills." But the skills here proposed to be tested are not questionable, according to currently available evidence, and their importance and criticality is well supported by theory and research. With regard to the claim that there exists "a vast array of possible subskills", I would comment that such an array could be conceived of only by enumerating reading behaviors at some microscopic level.

Point 6 argues that testing and reporting of subskills results may lead to a "fractionated, mechanistic" approach. Why any such result would occur is not made clear. There is no research evidence that I know of to the effect that knowledge of attainments in subskills would necessarily lead to a "fractionated, mechanistic approach," whatever that might be. I would suppose that such an approach might occur for any number of reasons other than the availability of "diagnostic" information. Various measures could be taken to guard against "mechanistic" approaches, but ignoring "diagnostic" information is not one of them.

The resolution's concluding statement, that reading instruction might be harmed because of potential misuse of test results, has no basis that I know of in research. Perfetti (1985, p. 241) remarks that "comparative studies continue to support the assumption that instruction that includes code teaching is advantageous for some things (reading) and not harmful for others." Test results of <u>any</u> kind--even the current NAEP

results--are always subject to misuse and misinterpretation, but it is certainly odd to claim that one is better off in the absence of information than when potentially useful information is available. In my view, reading instruction in the US could be improved if we knew more about levels of attainment of the reading skills that I have described above, because it has been demonstrated that these skills are potentially critical for the attainment of high levels of general reading skill.

In my notes about the controversy surrounding these subskill problems, it seems that there were some objections on the basis that word decoding and the other subskills were seen as "enabling" skills rather than as final outcomes. There was the feeling among certain IRA members that NAEP should not attempt to measure "enabling skills." I should point out, incidentally, that study and reference skills can be interpreted as "enabling" skills, and there seems to have been no objection to measurement of these skills in current and past NAEP assessments. But even if word decoding, for example, is viewed simply as an "enabling" skill, that is, a skill that enables or at least helps the reader to attain high levels of functional proficiency, this is no reason to reject it as a component of the NAEP assessment. On the contrary, in my view, assessment of this "enabling" skill should materially help in the interpretation of levels of overall reading proficiency attained. One might learn, for example, that many students between "rudimentary" and "basic" skill levels have on the average lower than expected proficiency in word decoding skills, and that some individuals even at much higher levels have

deficiencies in word decoding. Such information would be of use in making recommendations about the role of instruction in word decoding.

If a plan to include the above-mentioned subskills in future NAEP reading assessments is accepted, NAEP would have several options in carrying out the plan. Any of these options would of course require research and development, for new exercises would have to be developed and the results analyzed. (From the information I have at hand, I assume that currently available NAEP exercises would not be adequate for measurement of subskills.). One option would be to construct series of exercises for subskill measurements and then to report results on separate scales, more or less in the manner of the reporting on the Reading Proficiency Scale. Another option would be somehow to link subskill scales to points on the Reading Proficiency Scale, reporting, for example, the average subskill performances found at each level on the Reading Proficiency Scale, without separate reporting of subskill scales. Neither of these options assumes that any basic change would be made in the construction of exercises and the scaling of their results, for the Reading Proficiency Scale itself.

III. Other Issues

There has been some discussion of the possibility of conducting NAEP assessments at ages in addition to 9, 13, and 17. One proposal is to include some earlier age group such as 5 or 7. I find myself not in favor of such a proposal. I would assume

that too few children at age 5 would have made enough progress in reading to make assessment worthwhile. Possibly if an assessment at age 5 were to include various types of "reading readiness" exercises and a variety of background information, the results might have some value in evaluating the entry status of children as they begin reading programs, but I am dubious even of this. It should be remembered that the philosophy and structure of NAEP assessment preclude any consideration of longitudinal data for individual children.

My objection to a proposed assessment at age 7 appeals mainly to the fact that it would be too close to the age 9 assessment and I doubt that enough additional information would be gained to make it worthwhile from a cost-benefit standpoint.

On the other hand, I would be definitely in favor of an assessment of reading at the adult level--say, age 30, particularly if it were to include a truly representative sample of the U.S. population at that age. Such an asessment would resolve many of the debates on the true literacy level of the population.

IV. Summary and Recommendations

(1) The Reading Proficiency Scale established by NAEP should be accepted as it stands, but supplementary information should be furnished for it, if possible for the 1984 assessment data, and certainly for future assessments. This supplementary information should make reference to readability levels and grade placements

associated with the five designated levels of proficiency on the scale. In future assessments, information should be furnished for additional points on the scale below 150 and above 350.

(2) More could be done to make more prominent, in reports prepared for the public, the variability features of assessment data--even those from past assessments.

(3) For future assessments, consideration should be given to the construction of exercises that would permit measurements of proficiency in word decoding, vocabulary knowledge, and reading speed. Depending on the outcomes, results would be reported either in terms of separate scales for subskills or in terms of the levels of performance on them associated with selected points on the Reading Proficiency Scale.

(4) I see little need for an assessment at any age earlier than age 9, but I would strongly recommend conducting an assessment of the reading skills of a representative sample of the adult population, say at age 30.

References

Baron, J. (1977) Mechanisms for pronouncing words: Use and acquisition. In D. LaBerge & S. J. Samuels (Eds.), Studies in reading: Perception and comprehension. Hillsdale, NJ: Lawrence Erlbaum Associates.

Carroll, J. B. (1971) Development of native language skills beyond the early years. In C. Reed (Ed.), The learning of language (pp. 97-156). New York: Appleton-Century-Crofts.

Carroll, J. B. (1985) Defining abilities through the person characteristic function. In E. E. Roskam (Ed.), Measurement and personality assessment (pp. 121-131). Amsterdam: North Holland.

Carver, R. J. (1974) Improving reading comprehension: Measuring readability. [Place: Publisher?].

Chall, J. S. (1958) Readability: An appraisal of research and application. Columbus, OH: Ohio University Press.

Chall, J. S. (1983) Stages of reading development. New York: McGraw-Hill.

Curtis, M. E. (1980) Development of components of reading skill. Journal of Educational Psychology, 72, 656-669.

Educational Testing Service (1957). Sequential tests of educational progress: Technical report, reading, writing, listening, social studies, science, mathematics. Princeton, NJ: Author.

Frederiksen, J. R. (1982) A componential theory of reading skills and their interactions. In R. J. Sternberg (Ed.), Advances in the psychology of intelligence, Vol. 1 (pp. 125-180). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jackson, M. D., & McClelland, J. L. (1979) Processing determinants of reading speed. Journal of Experimental Psychology: General, 108, 151-181.

Klare, G. R. (1963) The measurement of readability. Ames, IA: Iowa State University Press.

National Assessment of Educational Progress. (1970) Reading objectives. Ann Arbor, MI: Author.

Perfetti, C. A. (1985) Reading ability. New York: Oxford.

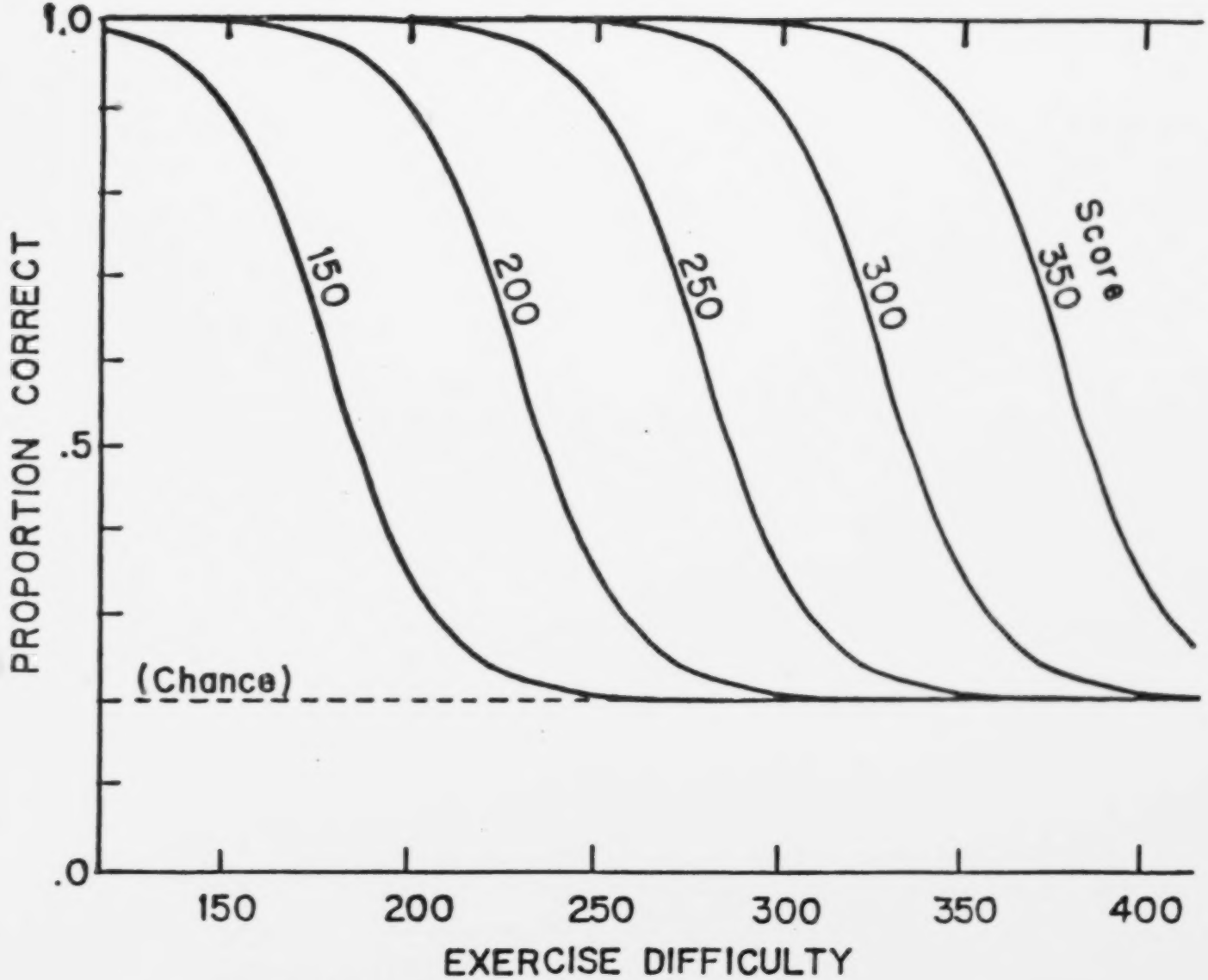


Fig. 1. Hypothetical person characteristic functions for students scoring at 150, 200, 250, 300, and 350 on the Reading Proficiency Scale.

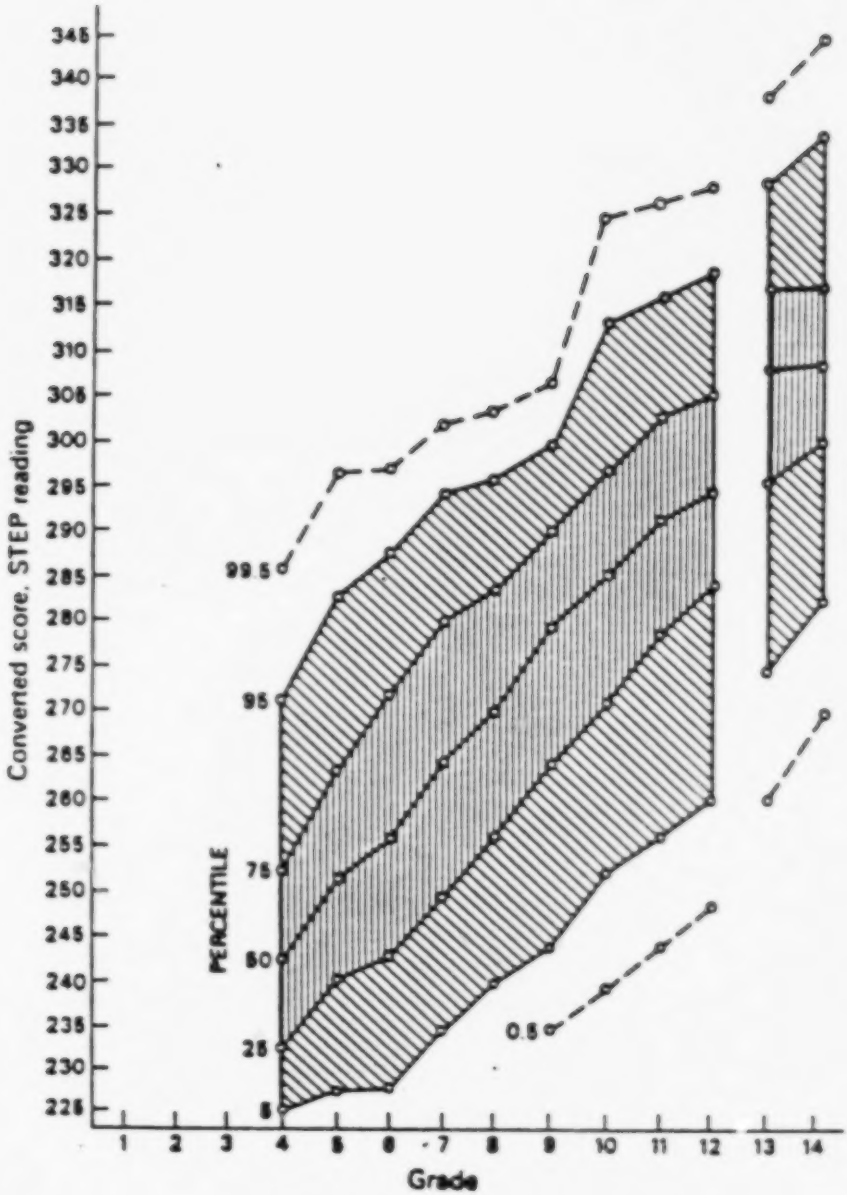


Fig. 2. Development and variations in reading ability at different grades as measured by the STP Reading Test (from Carroll, 1971, Fig. 3-1).

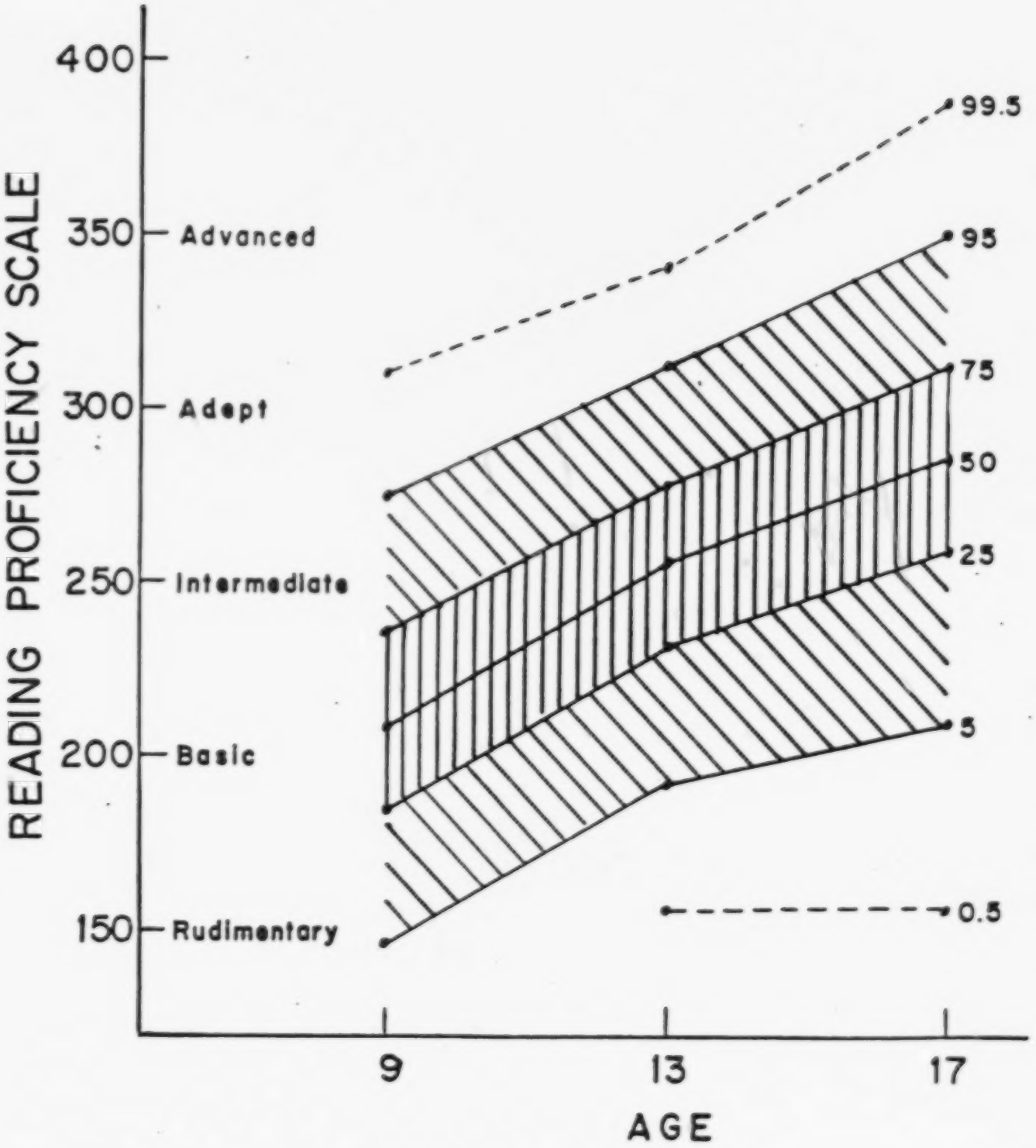


Fig. 3. NAEP Reading Assessment data, 1971.

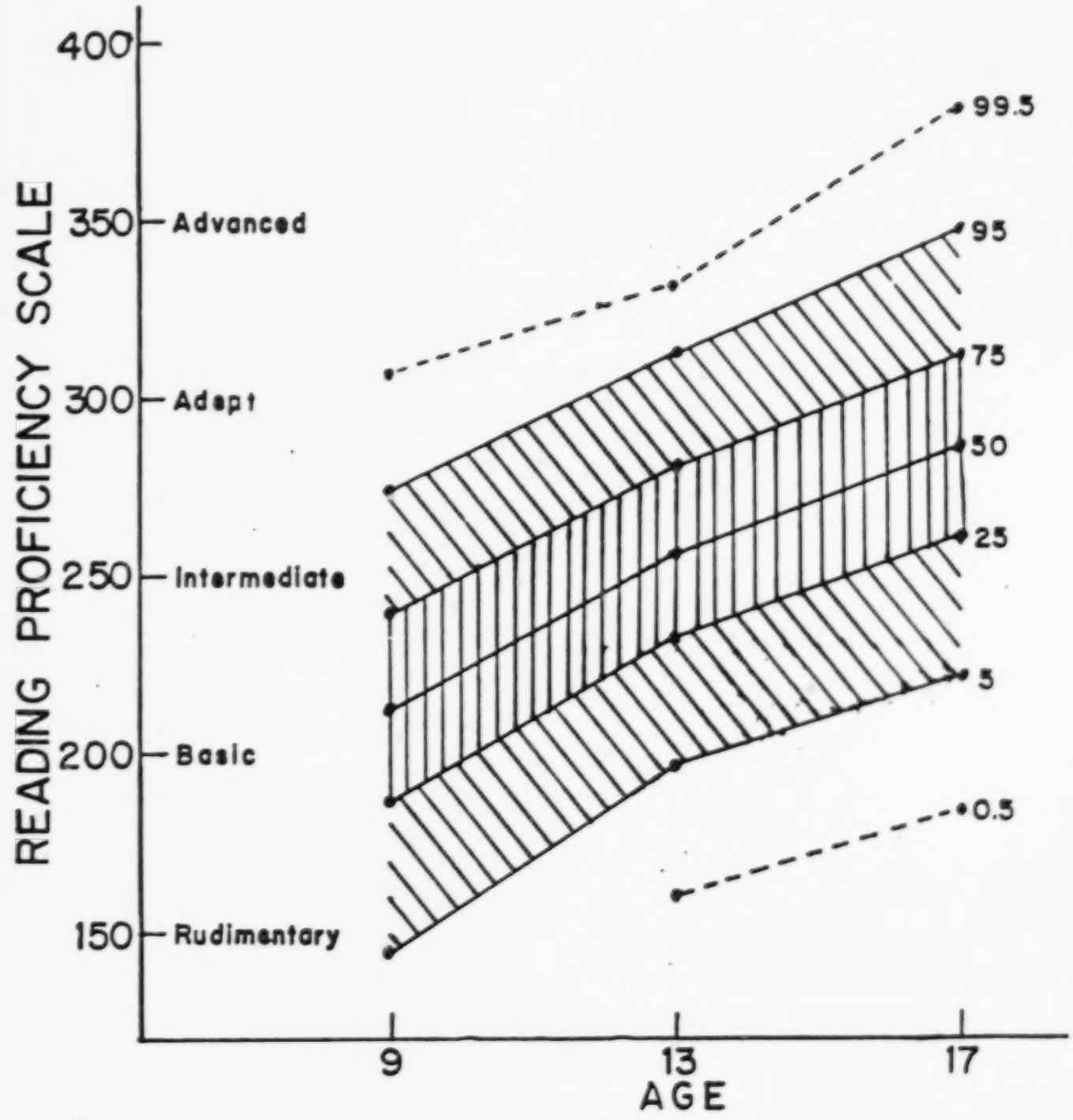


Fig. 4. NAEP Reading Assessment data, 1984.

Carroll analyses

Further analyses of data for disadvantaged/advantaged urban, 1971-1984

1. DISADVANTAGED URBAN.
Differences on RPS (1984 values minus 1971 values):

%ile:	0.5	5	25	50	75	95	99.5
Age 9	?	25	20	16	14	17	10
Age 13	6	8	7	5	4	2	1
Age 17	30	20	10	2	3	2	4

Age 9: Considerable, even marked improvement at all score levels
Age 13: Some improvement, more at lower score levels
Age 17: Most improvement is at score levels < 50%ile.

2. ADVANTAGED URBAN

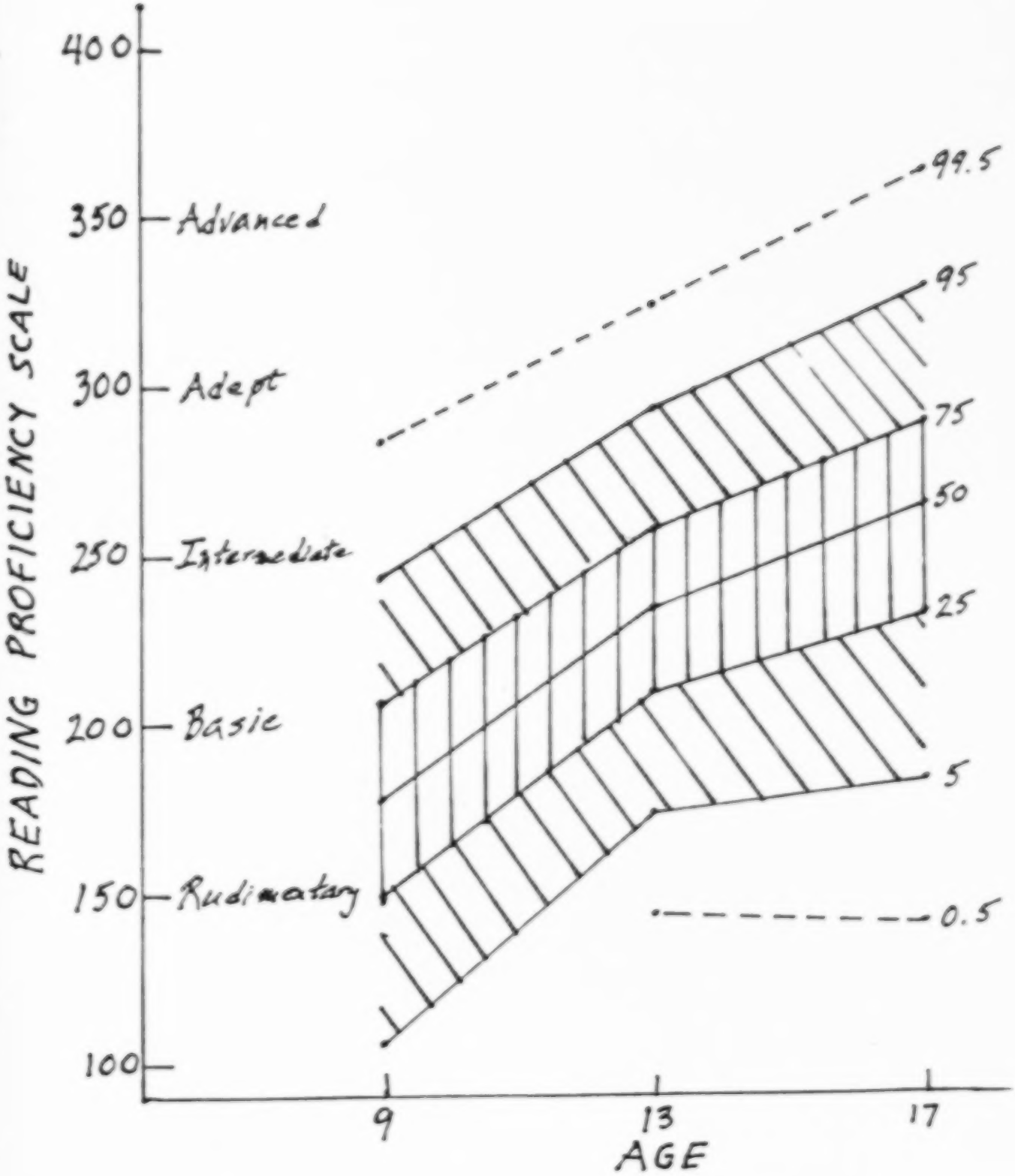
%ile:	0.5	5	25	50	75	95	99.5
Age 9	5	3	0	0	-1	-2	-2
Age 13	3	0	2	4	4	4	4
Age 17	-9	-4	-2	-2	-3	-3	-5

Age 9, 13, 17 : Hardly any marked difference at any age or score level. (Except slight decline for low ability age 17)
Thus, little change from 1971 to 1984. So, are advantaged urban already at best level to be expected considering wide variation in cognitive ability?

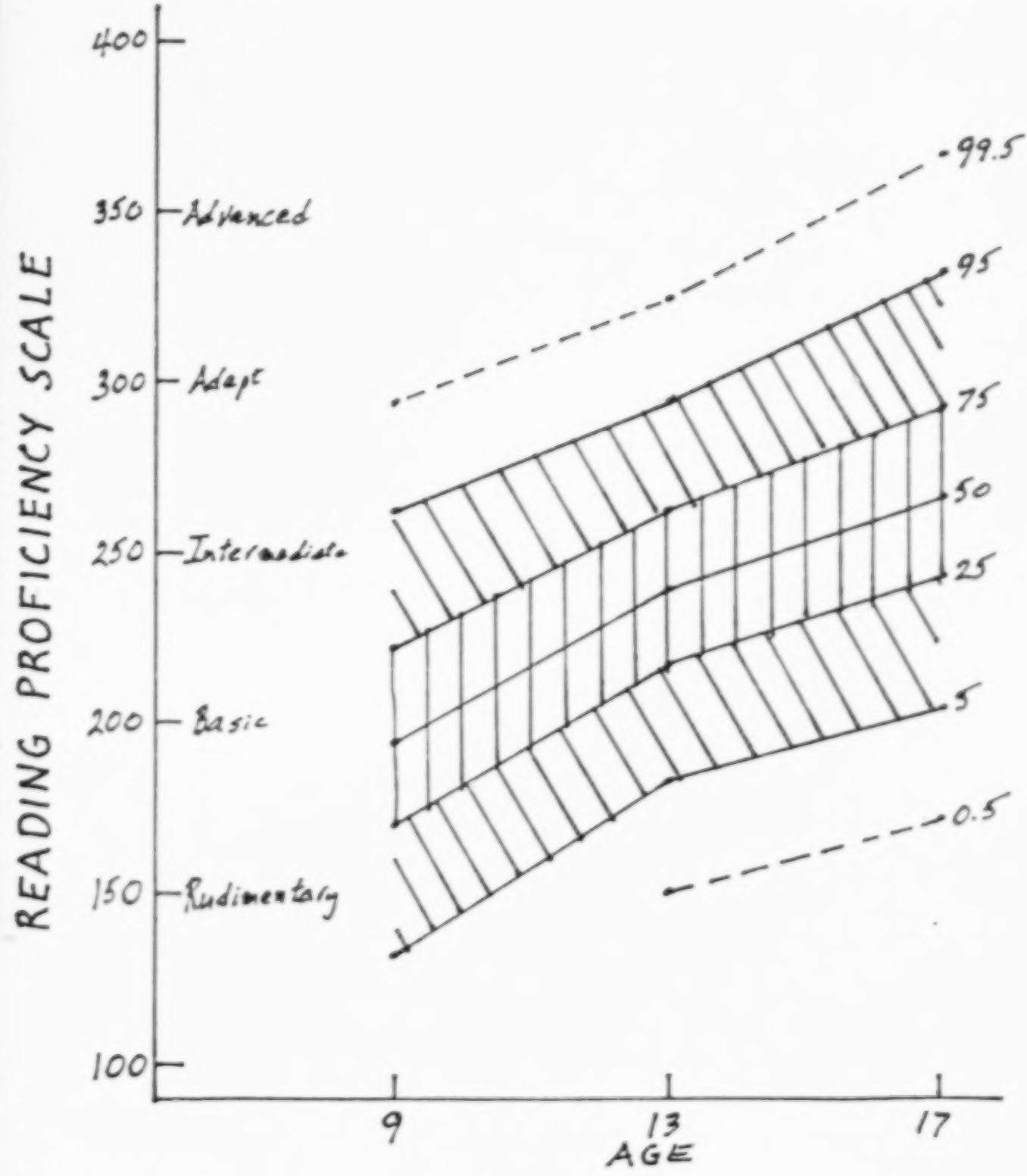
3. Comparisons of advantaged/disadvantaged
Differences on RPS (Advantaged value minus Disadvantaged value)

			1971 data				
%ile:	0.5	5	25	50	75	95	99.5
Age 9:	?	60	57	53	48	48	40
Age 13:	33	43	43	38	34	31	28
Age 17:	63	56	47	38	39	34	34
			1984 data				
Age 9	?	38	37	37	33	29	28
Age 13	30	35	38	37	34	33	31
Age 17	24	32	35	34	33	29	25

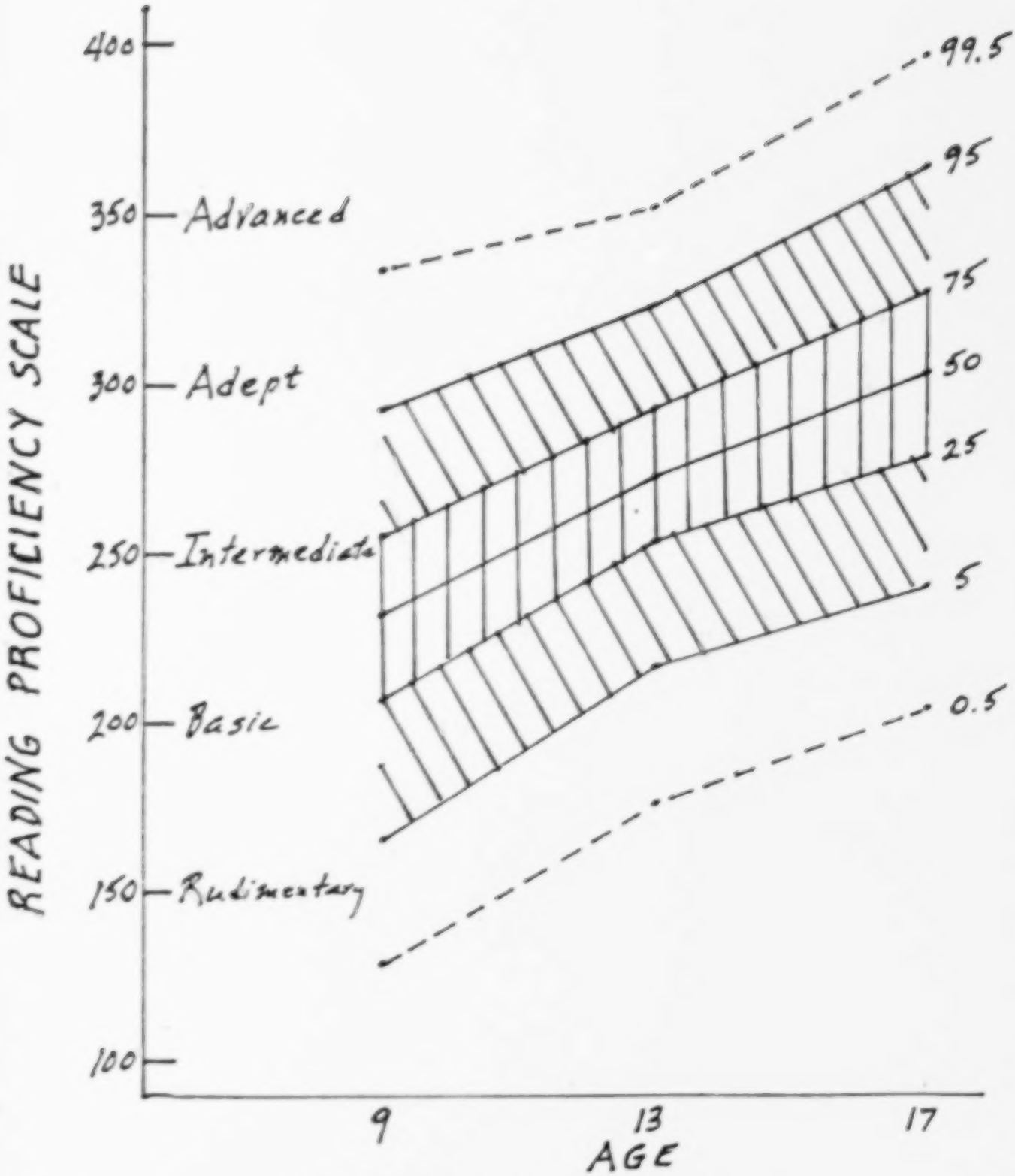
:: Differences between advantaged & disadvantaged continued to be large, but these differences narrowed considerably for low ability students at age 9 and age 17. Little change observed for Age 13 students (at any level)



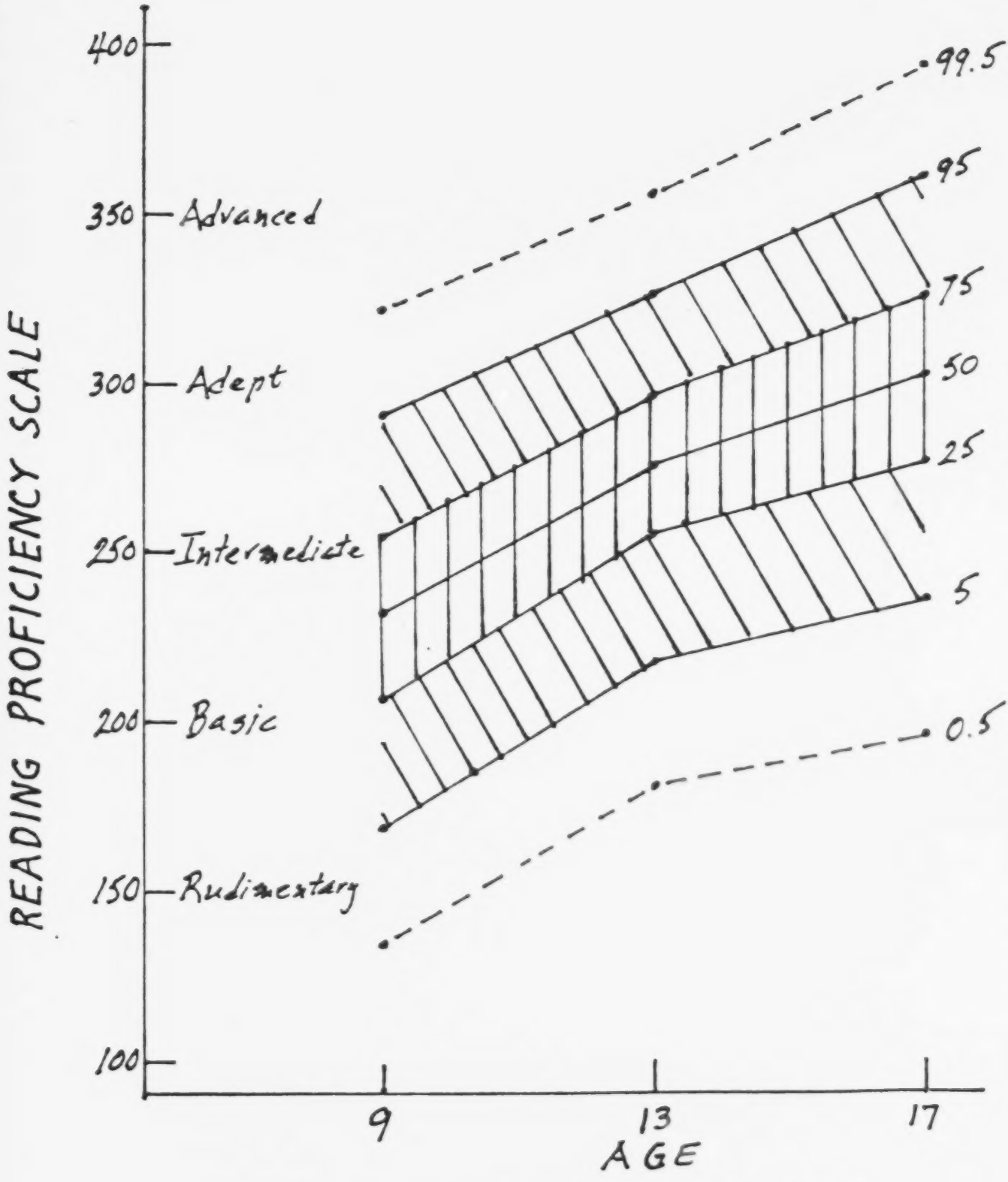
1971 DISADVANTAGED URBAN
(NAEP DATA)



1984 DISADVANTAGED URBAN



1971 ADVANTAGED URBAN



1984 ADVANTAGED URBAN